Description

~~Method for the production of a vertical MOS transistor~~.

With a view to ever-faster components with higher integration density, the ~~structure~~ structural sizes of integrated circuits are decreasing from generation to generation. This is also true ~~as~~ with regards CMOS technology. It is generally expected (see, for example, Roadmap of Semiconductor Technology, Solid State Technology 3, (1995)), that MOS transistors with a gate length of less than 100 nm will be used around the year 2010.

On the one hand, attempts have been made to scale modern CMOS technology in order to produce planar MOS transistors with such gate lengths (see, for example, A. Hori, H. Nakaoka, H. Umimoto, K. Yamashita, M. Takase, N. Shimizu, B. Mizuno, S. Odanaka, A 0.05 µm-CMOS with Ultra Shallow Source/Drain Junctions Fabricated by 5 keV Ion Implantation and Rapid Thermal Annealing, IEDM 1994, 485 and H. Hu, L. T. Su, Y. Yang, D. A. Antoniadis, H. I. Smith, Channel and Source/Drain Engineering in High-Performance sub-0.1 µm NMOSFETs using X-ray lithography, Symp. VLSI Technology, 17, (1994)). The production of such planar MOS transistors with channel lengths of less than 100 nm requires the use of electron beam lithography and has hitherto been possible only on a laboratory scale. The use of the electron beam lithography leads to a superproportional increase in ~~the~~ development costs.

In parallel with this, vertical transistors have been investigated with a view to producing shorter channel lengths (see, for example, L. Risch, W. H. Krautschneider, F. Hofmann, H. Schäfer, Vertical MOS Transistor with 70 nm channel length, ESSDERC 1995, pages 101 to 104). In this case, layer sequences are formed corresponding to the source, channel and drain, and are annularly surrounded by the gate dielectric and

gate electrode. In terms of their radiofrequency and logic properties, these vertical MOS transistors have to date been unsatisfactory in comparison with planar MOS transistors. This is attributed, on the one hand, to stray capacitances of the overlapping gate, and on the other hand to the formation of a parasitic bipolar transistor in the vertical layer sequence.

Summary Of The Invention

The object of the invention is therefore to provide a method for the production of a vertical MOS transistor, in which the radiofrequency and logic properties of the vertical MOS transistor are comparable with those of planar MOS transistors.

~~This object is achieved according to the invention by a method according to Claim 1. Further refinements of the invention are given by the other claims.~~

According to the invention, in the method, a mask with an opening is formed on a main surface of a semiconductor substrate, the main surface of the semiconductor substrate being exposed inside the opening. A layer sequence, which has one layer each for a lower source/drain region, a channel region and an upper source/drain region, is grown in this opening by selective epitaxy. During the growth of the layer sequence, facets are formed at the edge of the layer sequence, so that the thickness of the layers is smaller at the edge of the opening than in the middle. A gate dielectric and a gate electrode are formed on the edge of the layer sequence.

In the method, use is made of the discovery that, during the selective epitaxy, facets are formed at the edges of a mask since the growth rate in selective epitaxy is smaller at these edges. A study of the formation of facets in selective epitaxy is, for example, disclosed by L. Vescan, Radiative recombination in SiGe/Si dots...., Mater. Science and Eng. ~~B28, 1-8 (1994)~~ Vol 28 pp 1-8 1944.

This property of selective epitaxy is utilized to make the thickness of the layers of the edge of the layer sequence smaller than in the middle of the layer

sequence. The effect achieved by this is that the base width of the parasitic bipolar transistor formed in the middle of the layer sequence is greater than the channel width of the vertical MOS transistor formed at the edge of the layer sequence. The channel properties are therefore decoupled from the bulk properties in the layer sequence. Since the parasitic bipolar transistor has a greater base width than corresponds to the channel length of the vertical MOS transistor, the vertical MOS transistor dictates the properties of the structure.

The mask preferably consists of SiO_2 and/or Si_3N_4 at least at the surface. When a mask made of these materials is used, the thickness ratio between the middle and the edge of the layer sequence can be set, according to the growth conditions, between 2 and 3.

It is within the scope of the invention, during the formation of the mask, to form surface-wide a first insulating layer, a conductive layer and a second insulating layer, in which layers the opening is produced. The gate dielectric is formed on the exposed surface of the conductive layer before the selective epitaxy to form the layer sequence. The gate electrode is formed from the conductive layer. This method has the advantage that, during the production of the gate dielectric and the gate electrode, the side wall of the layer sequence is no longer subjected to an etching process.

In this case, the lower source/drain region is preferably grown such that it ends at the edge of the opening level with the first insulating layer. The channel region is grown such that it ends at the edge of the opening level with the conductive layer. This minimizes ~~the~~ stray capacitances of the gate electrode, which leads to a further improvement in ~~the~~ radiofrequency properties.

It is furthermore within the scope of the invention to form the mask from insulating material. After the layer sequence has been formed, the side wall

of the channel region is then exposed in such a way that the side wall of the lower source/drain region remains essentially covered by the insulating material of the mask. The gate dielectric and the gate electrode are subsequently formed on the exposed side wall of the channel region, the gate electrode preferably having its height matched to the height of the channel region. In this embodiment as well, the capacitances of the gate electrode are minimized, which leads to an improvement in the radiofrequency properties. The gate electrode is, for example, formed by depositing and structuring a conductive layer.

The mask of insulating material is in this case preferably formed from a first insulating layer and a second insulating layer. The first insulating layer is in this case arranged on the main surface of the substrate. The second insulating layer is arranged on the first insulating layer. The second insulating layer is etchable selectively with respect to the first insulating layer and with respect to the layer sequence. The lower source/drain region is in this case grown in a height such that it ends at the edge of the opening level with the first insulating layer. After the layer sequence has been grown, an opening, which annularly surrounds the channel region, is formed in the second insulating layer. After the gate dielectric has been formed, the opening is filled with a conductive layer. ~~Lastly,~~ Last, the gate electrode is formed by structuring the conductive layer, for example with the aid of planarizing steps.

In this case it is particularly advantageous to have the opening in the second insulating layer extend considerably beyond the layer sequence, at least on one side of the layer sequence. In this case, the opening has an extension at least on one side of the layer sequence. Island-like auxiliary structures made of the material of the second insulating layer are arranged in this extension. The opening consequently has a grid-like cross-section in the extension. The conductive

layer fills the opening in the extension as well. As a result, the gate electrode also at least partly has a grid-like cross-section. A contact hole, whose structural fineness may be substantially coarser than the structures in the opening, can subsequently be opened to the gate electrode in the extension. In this way, the contact hole can be dimensioned in such a way as to optimize electrical properties of the gate contact.

A further improvement in ~~the~~ radiofrequency properties by minimizing the stray capacitances is achieved in that the layer sequence is structured annularly and the annularly structured layer sequence is provided with an insulating filling. By removing the semiconductor material in the interior of the layer sequence, the formation of space-charge zones, which in turn cause stray capacitances, is suppressed.

~~The invention will be explained in more detail below with reference to illustrative embodiments which are represented in the figures.~~

Figure 1 shows a section through a semiconductor substrate with a terminal region and a mask.

Figure 2 shows the section through the semiconductor substrate after the formation of a layer sequence by selective epitaxy.

Figure 3 shows the section after formation of an opening which annularly surrounds the layer sequence and formation of a gate dielectric.

Figure 4 shows a plan view of Figure 3.

Figure 5 shows the section represented in Figure 3 after the opening has been filled with a conductive layer and a planarizing insulation layer has been produced.

Figure 6 shows the section after formation of a gate electrode by structuring the conductive layer.

Figure 7 shows the section after contact holes have been opened.

Figure 8 shows the section after formation of metal silicide terminal faces, a passivating layer and contacts.

Figure 9 shows the section through a semiconductor substrate with a terminal region and a mask.

Figure 10 shows the section after formation of a layer sequence by selective epitaxy.

Figure 11 shows the section after formation of an opening which annularly surrounds the layer sequence.

Figure 12 shows the section after formation of a gate electrode, a passivating layer and contacts.

Figure 13 shows a section through a semiconductor substrate with a terminal region and a mask which has a conductive layer on whose surface a gate dielectric is formed.

Figure 14 shows the section after formation of a layer sequence by selective epitaxy and deposition and planarization of an insulating layer.

Figure 15 shows the section after the insulating layer has been etched back and spacers have been formed on the side walls of the mask.

Figure 16 shows the section after the layer sequence has been structured annularly using the spacer as a mask, the surface of the terminal region being exposed.

Figure 17 shows the section after the annularly structured layer sequence has been provided with an insulating filling and after the formation of contacts.

The representations in the figures are not to scale.

In a monocrystalline silicon substrate 11, for example a monocrystalline silicon wafer or the monocrystalline silicon layer of an SOI substrate, a terminal region 12 is, in a first embodiment, formed by arsenic or phosphorous implantation at 5×10^{15} cm^{-2}, 40 keV and subsequent heat-treatment to activate the dopant (see Figure 1).

A mask 13 is subsequently formed on the substrate 11. To this end, a silicon nitride layer 131 is applied surface-wide in a thickness of, for example, 70 nm, and a silicon oxide layer 132 is applied thereon in a thickness of, for example, 500 nm. The silicon oxide layer 132 and the silicon nitride layer 131 are subsequently structured by anisotropic etching, an opening 130 being formed. The surface of the terminal region 12 is exposed inside the opening 130.

A layer sequence 14, which has a first layer 141 for a lower source/drain region, a second layer 142 for a channel region and a third layer 143 for an upper source/drain region, is grown inside the opening 130 by selective epitaxy (see Figure 2). The first layer 141 is, for example, grown from n-doped silicon with a dopant concentration of 5×10^{19} cm^{-3} in a layer thickness of 100 nm. The second layer 142 is, for example, grown from p-doped silicon with a dopant concentration of 10^{18} cm^{-3} in a layer thickness of 100 nm. The third layer 143 is grown from n-doped silicon with a dopant concentration of 5×10^{19} cm^{-3} in a layer thickness of 200 nm. The selective epitaxy is in this case controlled in such a way as to form facets on the edge of the opening 130. This means that, at the edge of the opening 130, the first layer 141, second layer 142 and the third layer 143 have a smaller layer thickness than in the middle of the opening 130. The specified layer thicknesses refer to the middle of the opening. The selective epitaxy is, for example, carried out using the following process gases: $Si_2H_2Cl_2$, B_2H_6, AsH_3, PH_3, HCl, H_2, in the temperature range of between 700 to 950°C and in the pressure range of between 5 to 20,000 Pa on silicon wafers with a [110] flat orientation. The first layer 141 is grown in such a way that its thickness at the edge of the opening 130 coincides approximately with the thickness of the silicon nitride layer 131.

Using a photolithographically produced mask (not shown), an opening 15 which exposes the side walls

of the layer sequence 14 is subsequently formed in the silicon oxide layer 132 (see Figure 3 and plan view in Figure 4). The surface of the silicon nitride layer 131 is exposed in the opening 15. Laterally with respect to the layer sequence 14, the opening 15 has an extension 150 in which island-like structures 132′ made of the material of the silicon oxide layer 132 are arranged (see Figure 4). The island-like structures 132′ are arranged as a matrix, so that the opening 15 has a grid-like cross-section in the extension 150.

The opening 15 laterally overlaps the layer sequence 14. Since the alignment in lithographic methods is more accurate than the minimum structural size, the distance between the layer sequence 14 and the structured silicon oxide layer 132 is less than a minimum structural size. When use is made of lithography with a minimum structural size of 0.6 µm and an alignment accuracy of 0.2 µm, the distance between the layer sequence 14 and the silicon oxide layer 132, or the island-like structures 132′, is, for example, 0.3 µm. The structural size of the island-like structures 132′ is in each case one minimum structural size, for example 0.6 µm.

An SiO_2 gate dielectric 16 is subsequently formed in a layer thickness of 3 to 5 nm by thermal oxidation on the exposed surface of the second layer 142 and the third layer 143.

A conductive layer 17 is subsequently deposited surface-wide. The thickness of the conductive layer 17 is set in such a way that the intermediate space between the layer sequence 14 and the silicon oxide layer 132 is filled. All materials appropriate for the gate electrode are suitable for the conductive layer 17, in particular doped polysilicon, metal silicide, metal. The conductive layer 17 is, for example, formed from n-doped polysilicon in a layer thickness of 400 nm (see Figure 5). A planarizing layer 18 is subsequently formed on the insulating layer 17, for example from a photoresist or a different spin-on material. The

surface of the conductive layer 17 is planarized, for example by planarizing etching or chemical/mechanical polishing. The conductive layer 17 is subsequently etched highly selectively with respect to SiO_2. In this case, a gate electrode 170 is formed from the conductive layer 17 (see Figure 6).

A further SiO_2 layer is subsequently applied surface-wide in a layer thickness of, for example, 70 nm and is structured with the aid of a photoresist mask 19. In this case the surface of the terminal region 17, of the gate electrode 170 and of the third layer 143 are partly exposed (see Figure 7).

By self-aligned siliciding, for example in a silicide process using titanium, silicide terminals 110 are formed on the exposed surface of the terminal region 12, the gate electrode 170 and the third layer 143 (see Figure 8). The silicide terminals 110 serve in each case to reduce the stray series resistances.

After surface-wide application of a passivating layer 111, for example of SiO_2, in which contact holes are opened to the silicide terminals 110, to the terminal region 12 and to the third layer 143 and to the gate electrode 170, contacts 112 to the terminal region 12, to the third layer 143, which forms the upper source/drain region, and to the gate electrode 170 are formed by forming a metal layer and structuring the metal layer. The contact hole to the gate electrode 170 cannot be seen in the section represented in Figure 8. It is in the extension 150 (compare Figure 4). By virtue of the grid-like structure of the gate electrode 170 in the extension 150 (compare Figure 4), it is possible to provide the contact hole to the gate electrode 170 with a larger cross-section than corresponds to the structural sizes of the gate electrode 170 in this area. The contact hole to the gate electrode 170 overlaps one or more of the island-like structures 132'.

In a substrate 21, for example a monocrystalline silicon waver or the monocrystalline

silicon layer of an SOI substrate, a terminal region 22 is, in a second embodiment, formed, for example by masked implantation and subsequent heat-treatment to anneal the defects due to the implantation. A mask 23 having an opening 230 in which the surface of the terminal region 22 is exposed, is subsequently formed on the surface of the substrate 21 (see Figure 9).

In order to form the mask 23, a terminal layer 231, a silicon nitride layer 232 and a silicon oxide layer 233 are applied to the substrate 21. The terminal region 231 is, for example, formed from heavily doped polysilicon in a layer thickness of 50 nm. All electrically conductive materials are suitable for the terminal layer 231, in particular doped polysilicon, silicide, metal. The silicon nitride layer 232 is applied in a layer thickness of 20 nm. The silicon oxide layer 233 is applied in a layer thickness of, for example, 500 nm.

By using a photolithographically produced mask (not shown), the terminal layer 231, the silicon nitride layer 232 and the silicon oxide layer are structured by anisotropic etching for example with CHF_3, O_2 (for nitride, oxide), HBr, Cl_2, He, O_2 (for polysilicon). The opening 230 is in this case formed. Silicon oxide spacers 234 are then formed on the side walls of the terminal layer 231, of the silicon nitride layer 232 and of the silicon oxide layer 233 which face the opening 230, by conformly depositing and anisotropically etching back a silicon oxide layer. The silicon oxide spacers have width of 10 nm (see Figure 9).

A layer sequence 24, which has a first layer 241 for a lower source/drain region, a second layer 242 for a channel region and a third layer 243 for an upper source/drain region is grown in the opening 230 by selective epitaxy (see Figure 10). The selective epitaxy is carried out while adhering to the following process conditions: process gas: SiH_2Cl_2, B_2H_6, AsH_3, PH_3, HCl, H_2, temperature range: 700°C to 950°C,

pressure range: 5 to 20,000 Pa. In this case, the first layer 241 is formed from n-doped silicon with a dopant concentration of 5×10^{19} cm^{-3} in a layer thickness of 100 nm. The second layer 242 is formed from p-doped silicon with a dopant concentration of 10^{18} cm^{-3} in a layer thickness of 100 nm. The third layer 243 is formed from n-doped silicon with a dopant concentration of 5×10^{19} cm^{-3} in a layer thickness of 200 nm. The specified thicknesses refer to the middle of the opening 230. The specified process parameters lead to the formation of facets on the edge of the opening 230, so that the layer thicknesses of the first layer 241, of the second layer 242 and of the third layer 243 are smaller there by a factor of about 2 to 3.

An opening 25 which annularly surrounds the layer sequence 24 is subsequently formed (see Figure 11). The side walls of the second layer 242 and of the third layer 243 are exposed in the opening 25. The opening 25 is etched using a photolithographically formed mask (not shown), the silicon nitride layer 232 serving as an etching stop. A residue of the silicon oxide spacer 234, which insulates the terminal layer 231 from the first layer 241, remains in the first layer 241. The terminal layer 231 is electrically connected to the terminal region 22.

A gate dielectric 26 is formed by thermal oxidation on the exposed surface of the second layer 242 and of the third layer 243. The gate dielectric 26 is formed from SiO_2 in a layer thickness of, for example, 3 to 5 nm. The MOS transistor is fabricated, like in the first illustrative embodiment, by forming a gate dielectrode 270 which fills the opening 25, by depositing and structuring a further SiO_2 layer 28, by forming silicide terminals 210 to the third layer 243, to the gate electrode 270 and to the terminal layer 231, by depositing a passivating layer 211 and by forming contacts 212 to the silicide terminals 210 which are arranged on the third layer 243, on the terminal layer 231 and on the gate electrode 270. The

contact 212 to the gate electrode is preferably provided laterally with respect to the layer sequence 24, as described with reference to the first illustrative embodiment.

In a substrate 31, for example a monocrystalline silicon wafer or the silicon layer of an SOI substrate, a terminal region 32 is formed in a third illustrative embodiment. The terminal region 32 is, for example, formed by As implantation at 5×10^{15} cm^{-2}, 40 keV and subsequent heat-treatment to anneal the defects due to implantation.

A mask 33 which has an opening 330 is then formed on the surface of the substrate 31. The surface of the terminal region 32 is partly exposed inside the opening 330 (see Figure 13).

In order to form the mask 33, a silicon nitride layer 331 in a thickness of 20 nm and a 50 nm thick first silicon oxide layer 332 are applied to the surface of the substrate 31. A conductive layer is applied thereon and is structured in such a way that it forms a gate electrode 370. The gate electrode 370 is formed from doped polysilicon in a layer thickness of 100 nm. A second silicon oxide layer 333 is applied thereon in a layer thickness of 600 nm and is planarized. The opening 330 is opened in the mask 33 by anisotropic etching using a photolithographically formed mask (not shown). The opening 330 has dimensions of, for example, 0.6×0.6 μm^2. This assumes lithography in which the mimimum structural size F = 0.6 µm and the alignment accuracy is at most 0.2 µm.

During the formation of the opening 330, etching is firstly carried out as far as the surface of the silicon nitride layer 331. An SiO_2 gate dielectric 36 is then formed in a layer thickness of from 3 to 10 nm by thermal oxidation on the exposed surface of the gate electrode 370. The silicon nitride layer 331 is then etched through selectively with respect to SiO_2 and with respect to silicon, the surface of the

terminal region 32 being partly exposed in the opening 330.

A layer sequence 34 is then grown in the opening 330 by selective epitaxy (see Figure 14). The layer sequence 34 has a first layer 341, a second layer 342 and a third layer 343. The first layer 341 is grown from n-doped silicon with a dopant concentration of 5×10^{19} cm^{-3} and a layer thickness of 150 nm. The second layer 342 forms a channel region and is grown from p-doped silicon with a dopant concentration of 10^{18} cm^{-3} in a layer thickness of 100 nm. The third layer 343 acts as an upper source/drain region and is grown in a layer thickness of 250 nm with a dopant concentration of 5×10^{19} cm^{-3} from n-doped silicon. The selective epitaxy is in this case controlled in such a way that the layer thicknesses at the edge of the opening 330 are less than in the middle of the opening 330. The specified layer thicknesses refer to the middle of the opening 330. The layer thicknesses are reduced at the edge of the opening 330 by a factor of about 2 to 3. The selective epitaxy is carried out while adhering to the following process parameters: process gas: SiH_2Cl_2, B_2H_6, AsH_3, PH_3, HCl, H_2, temperature range: 700°C to 950°C, pressure range: 5 to 20,000 Pa.

A 600 nm thick polysilicon layer 35 is subsequently applied and planarized with the aid of chemical/mechanical polishing selectively with respect to SiO_2. After the planarizing, the polysilicon layer 35 ends level with the second silicon oxide layer 333 (see Figure 14). The polysilicon layer 35 is preferably formed from n-doped polysilicon, so that it is electrically connected to the third layer 343.

The polysilicon layer 35 is then etched selectively with respect to SiO_2. In this case, a trench 37 is formed which has a depth of, for example, 300 nm (see Figure 15). The side walls of the second silicon oxide layer 333 are exposed in the trench 37.

On the side walls of the second silicon oxide layer 333 which are exposed in the trench 37, silicon

nitride spacers 38 are formed by conformly depositing a silicon nitride layer and anisotropically etching the silicon nitride layer back. The silicon nitride spacers 38 have a thickness of, for example, 50 nm.

The layer sequence 34 is then annularly structured in an anisotropic etching operation selectively with respect to silicon oxide and silicon nitride. The etching is continued until the surface of the terminal region 32 is exposed (see Figure 16). In this case, the silicon nitride spacers 38 act as mask. The free space formed inside the annularly structured layer sequence 34 is filled with an insulating filling 39. The insulating filling 39 is, for example, formed from SiO_2 by LPCVD deposition of a 400 nm thick SiO_2 layer and subsequent etching back. The silicon nitride spacers 38 are then selectively removed. Contact holes are thereby opened, with self-alignment, to the polysilicon layer 34 and therefore to the third layer 343, which acts as the upper source/drain region. Using a photoresist mask, contact holes, which extend as far as the terminal region 32 or the gate electrode 370, are then etched into the first silicon oxide layer 332 and the second silicon oxide layer 333, as well as into the silicon nitride layer 331 (see Figure 17). Contacts 312 to the gate electrode 370, to the polysilicon layer 35 and to the terminal region 32 are subsequently formed by applying and structuring a metal layer.